INNOVATING CROSS-DOMAIN SOLUTIONS TO DETECT EMERGING BIOLOGICAL THREATS

A Comprehensive Bioinformatics Pipeline for the Identification of Conserved Binding Sites Enables Broad-spectrum Biological Threat Detection and MCM Development

Cassie Bryan Draper Lab **John Julias** Draper Lab **Abby Aronica** Draper Lab **Samuel Barton** Draper Lab
Nel Bose Draper Lab **Brianna Fontak** Draper Lab **Alexander Hamme** Draper Lab **Michelle Karker** Draper Lab
Nick Last Draper Lab **Daniel Matera** Draper Lab **Jena Nawfel** Draper Lab **Ting Pang** Draper Lab

Identification of common targetable protein elements enables broad spectrum detection and medical countermeasure (MCM) development against large swaths of the biological threat space, including viruses, bacteria, fungi, and toxins. Protein regions that are conserved, or highly similar at the protein sequence and structure level, arise through functional selection as the pathogen evolves. In order to target these conserved regions – with a biosensor or therapeutic antibody, for example – they must have certain properties that make them capable of supporting a high affinity interaction with another molecule. Draper has developed a bioinformatics pipeline that uses both AI/ML and biophysics-based tools to identify conserved binding sites across large classes of proteins and organisms. This system allows for the targeting of not only current and past circulating pathogens and protein toxins but also unknown and emerging threats that are not yet characterized.

Developed as part of IARPA's B24IC program to facilitate viral proteome-wide epitope identification for uncharacterized antibodies, the C2PEP Pipeline uses structural information to identify conserved binding sites across large sets of proteins. The bioinformatics pipeline is modular to be easily customizable to various applications and is fully automated requiring only an input virus family name and protein name. It consists of five modules – Input Preparation, Epitope Prediction, Conservation Analysis, Scoring, and Database Generation. The pipeline integrates a suite of software tools to model conformational epitope probability and structural similarity and output a conserved epitope database. These tools rely upon experimental structure data, and we have also begun development of a Modeling module which will enable automatic generation of high-quality structural models in the absence of missing experimental data to improve our predictions. Performance of the C2PEP prediction pipeline has been assessed using a benchmark data set of 12 known conserved epitopes across 7 different proteins from 4 families of viruses.

By integrating both traditional biophysical methods of protein structure analysis with next-generation AI/ML-based tools, we maximize the accuracy of our predictions while expanding the general use and application space of the pipeline beyond the limitations of a single tool. The pipeline has been developed with flexibility and broad utility in mind; modules are able to be added or replaced and parameters adjusted easily without retooling the entire pipeline to customize it to a different target application. Conserved binding site predictions are output in an actionable format that can immediately inform or be integrated with a design pipeline for broad-spectrum MCM or biosensor development. The C2PEP pipeline is a robust generalized framework adaptable to applications in surveillance, detection, and MCM development against known and emerging biological threats.

The research presented is supported by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), under B24IC program via contract N66001-23-C-4501. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the views or official policies, either expressed or implied, of ODNI, IARPA, or the U.S. Government.